COMPUTER-BASED STORAGE AND RETRIEVAL OF PALAEONTOLOGICAL DATA AT THE SEDGWICK MUSEUM, CAMBRIDGE, ENGLAND

by DAVID PRICE

ABSTRACT. The computer-based data handling system at the Sedgwick Museum is a specialized application of the GOS program package, geared to the requirements of a large palaeontological collection. Rigorous analysis and careful structuring of palaeontological data facilitate the automatic production, by machine processing, of hard-copy catalogues, labels, and a variety of indexes. The system goes beyond the capabilities of the standard GOS package in that a number of extra programs are geared to direct access information retrieval. The Information Retrieval System incorporates a solution to the problem of variant forms of single catalogue terms and can be used to answer rapidly many queries about the collections which formerly would either have been virtually impossible or have required extensive physical searches. Those aspects of the Sedgwick System of interest and potential use to palaeontologists and biostratigraphers are described.

MANY museums are using or beginning to use computer-based methods for cataloguing and the production of collection indexes (see, for instance, Roberts and Light 1980). The computer-based procedures described in this paper will be of particular interest to palaeontologists in that they are geared to the documentation requirements of the Sedgwick Museum, whose large palaeontological collections are of international scope and contain a high proportion of type, figured, and cited specimens; many represent classic pieces of research.

HISTORICAL DEVELOPMENT

The long history of Cambridge as a centre of palaeontological and biostratigraphic research ensured that the value of well-organized and accessible collections was appreciated at an early date (see Rickards 1979). Such appreciation led to the development, particularly during the pioneering regime of A. G. Brighton (Curator, 1931–1968), of a curatorial system whose outstanding characteristic was its exceptionally high standard of collection documentation. Specimens were comprehensively labelled, fully and systematically catalogued, and taxonomically indexed; there was also a supporting Curator's library of annotated scientific periodicals, monographs, and reprints, together with collector's original catalogues, notebooks, field-slips, and museum correspondence. This Sedgwick manual system has been described elsewhere (Price 1981; see also Orna and Pettitt 1980, pp. 152–153).

The Sedgwick Museum became a centre for the early development of ideas on computer-based data handling in museums, through the work of J. L. Cutbill, from the mid-1960's onwards (Cutbill 1971; Cutbill *et al.* 1973). Many characteristics of the Sedgwick manual system had an important influence on those ideas. But the 'computerization' of the Sedgwick Museum came to be only a partial aim of the several Sedgwick-based projects through which Cutbill's work developed. Those projects increased in scope to embrace a multidisciplinary approach to the analysis and recording of museum data, and the development of a computer program package for handling museum data which was equally flexible. While this broader approach led in 1977 to the setting up of the Museums Documentation Association, and the subsequent release by MDA of the GOS program package, the move of the final Sedgwick-based project team to their new MDA headquarters (Duxford,

[Palaeontology, Vol. 27, Part 2, 1984, pp. 393–405.]

Cambridgeshire) took place before the Sedgwick manual catalogue had been entirely transcribed to machine-readable computer files and indeed before all the necessary procedures for maintaining and updating a computerized catalogue had been fully developed and tested.

The gap between 1977 and the revival of the Sedgwick computerization project in 1981 only served to emphasize differences between the MDA's 'all-embracing' approach, with its broadly conceived data standard, and the rather specialized documentation procedures used in earlier phases of work at the Sedgwick Museum. Differences of approach have further increased during the completion and refinement of the Sedgwick System, because the final phase had the sole aim of deriving a working system specific to the needs of the museum. The final Sedgwick System is based on the GOS program package (with extra programs related to the Information Retrieval System described below), but is a specialized application of that package to a large palaeontological collection. Porter (1982) has given a more technical overview of the general capabilities of GOS. At the time of writing, the Sedgwick System's computer catalogue file contains details of 451 000 specimens, representing approximately 10 000 man-hours of typing and checking catalogue data.

It is not appropriate to describe here the Sedgwick System in its entirety. It is a complete working system covering all aspects of museum documentation, including the automatic production of labels and the recording of specimen loans. This paper attempts merely to describe those characteristics and capabilities of the Sedgwick System which will be of interest and potential use to palaeontologists and biostratigraphers.

GENERAL STRUCTURE OF DATA

Palaeontological data is stored in a precisely structured, machine-readable form within the computer file which now constitutes the Sedgwick Museum catalogue. To enable manipulation by the various programs ('processors') of the GOS package, each record is broken down into a large number of discrete data-categories or *fields*; each field is labelled or *tagged* so that it can be identified by the programs operating on it. The nature and relationships of all the fields considered necessary to contain the data in any Sedgwick Museum catalogue record are described by the *SM Format*, a complex hierarchical arrangement of fields derived from a rigorous analysis of existing manual catalogue entries, together with a consideration of what other kinds of data it might be useful to record for each specimen. The SM Format is described in the Appendix.

RETRIEVAL OF DATA

Indexes

The original purpose of rigorously analysing the data in each record into tagged fields (in the way outlined in the Appendix) was to enable the generation of collection indexes on a very large number of keyword terms. The possible use of any of the tagged keyword terms in the format to sort and order the records permits the construction of complex taxonomic indexes, donor indexes, collector indexes, locality indexes, stratigraphic indexes, bibliographic indexes, and so on. The generation and use of such indexes for data retrieval was indeed central to all thinking about data-handling in the Sedgwick Museum up until the most recent phase of 'computerization'. The Sedgwick System still retains the ability to produce all these indexes, but their use in practice has involved a number of difficulties.

Most indexes have proved very unwieldy in use, particularly so when two or more large indexes are used in conjunction. Problems of this kind can be reduced somewhat by using the great flexibility of the GOS package to produce 'multiple' indexes. An example index is shown as text-fig. 1 arranged primarily geographically, but including also basic stratigraphy and taxonomic names. Another way of easing the problems of handling large indexes is to reduce their bulk by producing them as microform output (COM), in our case as 127 × 76 mm microfiche at a 42 × reduction. Even so, difficulties with hard copy indexes remain.

It should be emphasized at this point that much of the machine-readable catalogue at the Sedgwick Museum was simply transcribed, word for word, from a manual catalogue which had grown

British Somaliland, Biyo Gora.
Eocene,
(Fish) Cichlid (indet.); (Sect. 1). C.80131
(Fish) Ogygiochromis sp.; (Sect. 1). C.80130
Lower Eocene,
(Fish) Pycnodont; (2628-2646 m below), (strata section i), 10 d 22' N., 45 d 12' E., loc: sigma 35. C.73153
(Fish) Sparoid; (2755-2781 m below top - strata section i), 10 d 22' N., 45 d 12' E., loc: sigma 42. C.73157-73158
Middle Eocene,
(Fish) Odontaspis sp.; (1960 m. below top - strata section i), loc: sigma 25. C.73155
Tertiary,
(Fish) Aplocheilus sp. nov.; (922-923 m. below top of strata section i), loc: sigma 17. C.76120-76132
(970-976 m. below top of strata section i), loc: sigma 18. C.76133-76134
(1179-1184 m. below top of strata section i), loc: phi 21. C.76135
(Fish) Tilapia? sp. nov.; (111-116 m. below top of strata section i), loc: phi 13. C.76087
(210-212 m. and 271-285 m. below top of strata section i), loc: phi 14. C.76088-76096
(strata section i), loc: phi 14. C.76097-76103
(270-274 m. below top of strata section i), loc: sigma 15. C.76104
(329-330 m. and 420-422 m. below top of strata section i), loc: phi 15. C.76105
(583-589 m. below top of strata section i), loc: phi 151. C.76106-76108
(922-938 m. below top of strata section i), loc: sigma 17. C.76109-76115
(970-976 m. below top of strata section i), loc: sigma 18. C.76116
(1179-1184 m. below top of strata section i), loc: phi 21. C.76117
(1199-1203 m. below top of strata section i), loc: phi 23. C.76118
(1326-1328 m. below top of strata section i), loc: phi 149. C.76119
(1179-1184 m. below top of strata section i), loc: phi 21. C.76136-76157
(1199-1203 m. below top of strata section i), loc: phi 23. C.76158-76165

British Somaliland, Biyo Gora, Daban Corner.
Eocene,
(Fish) Cichlid (indet.); (Sect. 1). C.80129

British Somaliland, Rhabka.
Eocene,
(Fish) (section), 10 d 10' N., 45 d 19' E., loc: sigma 14. C.73159-73161

Buckinghamshire, Brickhill.
Lower Greensand,
(Fish) Asteracanthus sp.; B.26675-26676
B.58620
(Fish) "Edaphodon" sp.; B.26759
(Fish) Ischyodus townsendii (Buckland); B.26765-26766
(Fish) Lepidotus maximus Wagner; B.26549-26553
(Fish) Otothus sp.; B.26790-26792
(Fish) Oxyrhina; B.26639
B.26648
(Fish) Pycnodus sp.; B.26637
(Fish) Pycnodus couloni Agassiz; B.26618-26619
(Fish) Sphenonchus sp.; B.26716
(Fish) Strophodus sp.; B.26595

Buckinghamshire, Pitstone, Upper Icknield Way.
Turonian,
(Fish) Scapanorhynchus subulatus (Agassiz); (pit on S. side), (600 yds. E.S.E. of church), grid ref: 42/946147. B.91737 4.20

TEXT-FIG. 1. Part of 'Fish' Locality Index.

gradually over almost fifty years. Data quality is thus rather variable. Over the full range of the catalogue, locality names may have several different spellings or differ in orthography. Geographical information for a single locality may be given in different records in different hierarchical sequences, in one case say 'farm name — village — town — county', in others 'farm name — town — county', or just 'village — county'. Such variation can cause a single locality to appear in many different places on a single index. The same is true for stratigraphic horizons.

Because of all these difficulties with indexes, and to obviate the need for reprinting large numbers of indexes at each updating of the catalogue, the Sedgwick Museum has now reduced the number of routinely used indexes to just one: a taxonomically based index divided up into sections on the basis of convenient suprageneric groups. An example of the layout and contents of part of the 'Fish' section of this index is shown in text-fig. 2. This taxonomic index in the form of microfiche (currently 90 fiche), together with a fiche version of the entire catalogue in alphanumeric order (117 fiche), are the only fixed hard-copy documents which play any important part in the Sedgwick System. The need for a variety of other indexes has been obviated by the use of an Information Retrieval System (IR System), which is now central to all Sedgwick Museum procedures.

Acanthodes sp.

(bone-bed - conglomeratic sandstone - Bed 2), Downtonian (base); Lower Wolton Farm, Woolhope Inlier, Herefordshire.
Listed, Gardiner, 1927, Q.J.G.S., lxxxiii, pp.517, 527 p.527. A.45356

Coal Measures, Carboniferous. E.3969

Coal Measures, Carboniferous; Newcastle, Northumberland.
Identified, Traquair, R.H. E.3970

Lancashire, 40 yds Mine, Coal Measures, Carboniferous; Bacup.
styliform, bone. E.3971

Acanthodes mitchelli Edgerton

Mesacanthus mitchelli (Egerton)

Old Red Sandstone, Devonian; Reswallie, Forfar. H.4465-4467

Old Red Sandstone, Devonian; Forfar. H.4468

Acanthodes nitides A.S.Woodward

Cement Stone Group, Calciferous Sandstone Series, Carboniferous; Esk R., Glencartholm, Langholm, Dumfries, Scotland.
Topotype, E.4965

Acanthodes wardi Egerton

Coal Measures, Carboniferous; Longton, Staffordshire. E.3950

Coal Measures, Knowles Ironstone, Carboniferous; Fenton, Staffordshire.
Identified, Traquair, R.H., spine. E.3951-3954

Coal Measures, Carboniferous; Longton, Staffordshire.
Identified, Traquair, R.H., pectoral, spine. E.3955

Coal Measures, Better Bed, Carboniferous; Low Moor, Yorkshire.
Identified, Traquair, R.H., spine. E.3956-3959

Coal Measures, 40 yds Mine, Carboniferous; Bacup, Lancashire.
Identified, Traquair, R.H., spine. E.3960-3961
Identified, Traquair, R.H., pectoral, spine. E.3962

Coal Measures, Carboniferous; Longton, Staffordshire.
trunk. E.3963-3965

TEXT-FIG. 2. Part of 'Fish' Taxonomic Index.

Information Retrieval System

The entire catalogue can be searched rapidly on-line using the IR System, on the basis of predetermined search criteria which may be very complex. The IR System responds initially by giving the number of *specimens* found in all records conforming to the search criteria. It can then be asked to list all such specimens by number, or a 'job' can be submitted to the computer to retrieve the actual records from a magnetic disc version of the catalogue. The job may take fifteen minutes or so to run.

Each specimen in the catalogue is indexed by a set of *terms* which are generated automatically from the GOS record describing the specimen. A term has the form of an initial upper case letter followed by a series of up to 23 lower case letters, digits, commas or full stops, but with all other characters, including spaces, discarded e.g. 'Ajones,o.t.1948', 'Roxfordclay'. The initial letter indicates the nature of the term: 'A', for instance, indicating authorship. The list below explains the significance of each upper case initial and shows also from which field or fields of a record each term is derived (field-tags, such as *a1 or *gn, relate to the SM Format given in the Appendix).

Term category	Example	Term derived from (see Appendix)
A = authorship	'Abarker,r.w.1927'	first *a1 and *ryear in *doc in each *re
D = donor	'Dstricklandcoll.'	each *ps1 in *oh
F = function word	'Fholotype'	each *f1 containing 'fig'd' or ending in 'type'
G = group	'Gtrilobite'	each *gn
K = keyword	'Kcranidium'	individual words in each *kw1
L = locality	'Lrobinhoodsbay'	each *loc1
N = informal name	'Ngoniatite'	each *taxs with no formal taxonomic name components
O = lithology	'Oblackshale'	each *lith1
P = preservation	'Preplacement'	each *pres1
Q = age	'Qjurassic'	each *age1
R = rock	'Rcornbrash'	each *rk1
S = store	'Sxxx.n.39'	each level (between stops) in each *store1
T = genus	'Thildoceras'	each *gen
U = species	'Ublumenbachii'	each *spec

In addition every catalogued specimen is indexed as 'I'. The way in which such index terms are generated from an actual record is illustrated by an example in the Appendix.

Within the system each term has a *term number*, *n*, which merely indicates its position within the current index, and a *frequency*, *f*, which is the number of specimens indexed by that term. Terms are usually printed in *n* — term — *f* order, with *f* in brackets, e.g.

593 = Gforaminiferan (5424)
2027 = Lperu (1921)

The simplest possible retrieval request is based on a single term. The query a-q <'Gammonoid'> asks what ammonoids there are in the collection. More complex queries combine terms together according to the rules of Boolean logic, using the operators '&' (= logical and), '|' (= logical or), and '—' (= logical subtraction). Further, actual terms may be replaced in queries by the term number, *n*, or by a specimen identity number, or by a range of such numbers. The query a-q <'Gammonoid' & 'Qoxfordian'—[F.1-23000]> asks for all British ammonoids of Oxfordian age (British since all 'foreign' Mesozoic specimens, of which there are just over 22,000, have specimen numbers prefixed with 'F.'). Of course, an actual retrieval request would be more complex than this, since not all records for ammonoids of Oxfordian age would necessarily contain 'Oxfordian' as an age term, particularly if they had been transcribed from the earlier part of the manual catalogue. In practice it would be necessary to retrieve also on rock terms such as 'Roxfordclay', 'Rampthillclay', 'Rwestwaltonbeds', 'Rbrorasandstone', etc. Similar problems arise when locality names appear in different guises in different parts of the catalogue.

The problem of variant forms of a single term. This is another aspect of the problem discussed earlier in relation to the use of hard-copy indexes; only in the case of index terms in the IR System there is a solution. The IR System has a facility for listing all index terms which show similarity to a given term. Each term in the term index is broken into fragments of four characters and separately indexed under all of these fragments, e.g. 'Cambridge' generates the fragments 'camb', 'ambr', 'mbri', 'brid', 'ridg', and 'idge'. The measure of similarity between terms is based on the number of fragments they have in common. A list can be requested which will give similar terms in decreasing order of similarity up to any number specified by the user. The list may either be restricted to terms of a particular category (e.g. age terms, genus names) or may include all terms irrespective of category.

The usefulness of this 'similar term' facility can be illustrated by an actual example. An IR query asking what specimens in the collections were donated by F. R. Cowper-Reed might be expected to take the form a-q <'Dcowperreed,f.r.'>. In fact, the submission of such a query should be preceded by a request for similar terms. If the ten most similar donor terms are requested by a-simt <Dcowperreed, f.r.>, the system responds with:

(cfterms)

Dcowperreed,f.r.:	3271 = Dcowperreed,f.r.	(804)	3269 = Dcowperreed	(22)
	3272 = Dcowperreed,f.r.coll.	(1)	4752 = Dreed,f.r.c.	(1225)
	3270 = Dcowperreed,f.	(3)	4753 = Dreed,f.r.c.coll.	(61)
	3254 = Dcooperreed,f.r.	(3)	4751 = Dreed,f.c.r.	(1)
	4754 = Dreed,f.r.cowper	(48)	5484 = Dwoodward,f.r.	(1)

The first nine of the above terms are indeed variations on F. R. Cowper-Reed ('cooperreed' is a misspelling; 'coll' is an abbreviation for collection). Using term numbers, the original query could now be replaced with a-q <3271 | 3272 | 3270 | 3254 | 4754 | 3269 | 4752 | 4753 | 4751>. The original query would have yielded only 804 specimens, but the more complex query now yields 2168.

The IR System can be used to answer many enquiries about the collections which would have been virtually impossible to answer using the manual system. Enquiries which formerly could only have been answered by physically searching the collections (exploiting a lay-out which reflects both stratigraphy and geography) can now be answered in a minute fraction of the time such a physical search would have taken—and much less fallibly.

DISCUSSION

Major benefits of 'computerization'

As a result of adopting computer-based data handling techniques, the Sedgwick Museum has acquired a wide range of new capabilities. Many of these are of strictly curatorial interest in that they are aimed at facilitating and improving management of the collections. Even here though external users will derive some benefit; for example, from rapid automatic processing of loans. Most of the main benefits to external users, however, are implicit in the preceding description of data retrieval, particularly in the illustration of the great variety of queries about the collections which can now be rapidly answered using the IR System. One major use of the IR System has been to provide palaeontologists with what might be termed 'specialist catalogues': retrieving and printing-out for instance the records of vertebrates from the Oxford Clay, Ordovician ostracodes, British Liassic, Callovian and Kimmeridgian corals, or Wealden reptiles from Brook on the Isle of Wight (all actual examples).

The complete museum catalogue and the taxonomic index are both on microfiche, and can therefore be very quickly and cheaply reproduced (in whole or in part) and readily distributed. Diazo copies from the COM fiche originals at present cost between 12p and 14p, depending on the numbers involved. At such a rate, a copy of the entire catalogue costs approximately £16, the taxonomic index approximately £13. Each section of the taxonomic index (e.g. 'trilobites', 'graptolites', 'bivalves') is usually only a few fiches in length.

Specimen security and locality conservation

The possibility of easy reproduction and distribution of catalogues and indexes raises problems of data security. At present the main concern must be with data about specimen storage. Data 'in the computer' is reasonably secure because it is only accessible to those at once familiar with the general working of the Cambridge Data Network, with the general working of GOS, and with the esoteric command syntax for the Sedgwick System; and who, in addition, are in possession of (or able to obtain) the passwords which safeguard access. Such data is also safe from loss because there are duplicated magnetic tape versions of the catalogue stored separately from the primary version and, of course, hard-copy fall-back in the form of fiche.

It is hard-copy output which potentially leads to the greatest security risk. Internal 'working' Museum hard-copy needs to carry storage locations; the present policy is to ensure that all hard-copy for external use is produced from GOS files which have had the *store field stripped out. It would be possible to strip out also detailed locality information, including grid-references or latitude and longitude. At present we see little point in doing this since the information is usually readily available in scientific papers and monographs, and because catalogue information is only distributed to bona fide research workers.

Future developments

Direct access to the museum database is restricted at present to the Curators who alone decide what information to release and to whom. It may be possible in the future to permit some form of access to external users at other computer sites linked to the Cambridge Network, or through a dial-up link. Such access would be via the IR System and, we envisage, would be restricted to the IR System indexes, thus enabling external users to make full use of the fiche versions of the catalogue and indexes while still safeguarding access to museum storage data (and any other data deemed 'sensitive' in the GOS database). Ideally any such arrangements would include reciprocal arrangements with other museums.

The Sedgwick Museum has acquired a complete, working, computer-based documentation system earlier than many other museums. It is now open to palaeontologists and palaeontological curators elsewhere to appraise the system in relation to their own requirements. It is our strong hope at this early stage that it might find wider application. Thoughts of a unified palaeontological database, involving the collections of many major institutions, may seem Utopian at present but we believe that in our work at the Sedgwick there lies one opportunity for such a future. Opposed to this opportunity is the very real danger that soon each institution will develop its own computer-based system, with its own data standard, and that mapping data from one system to another will become virtually impossible. In the field of computer-based museum documentation as a whole in Britain, it is already apparent that various factors (not least political and economic ones) have conspired to force such a pluralistic future. Perhaps for the major palaeontological collections this can yet be avoided.

Acknowledgements. My enormous indebtedness to Martin Porter, prime architect of the computer-based system described in this paper, will be obvious. He, Dr. W. D. I. Rolfe, and Mr. R. H. Hughes kindly read earlier typescripts and made useful suggestions for improvements.

REFERENCES

CUTBILL, J. L. 1973. *Computer filing systems for museums and research.* Sedgwick Museum, Cambridge.

—— HALLEN, A. J. and LEWIS, G. D. 1971. A format for the machine exchange of museum data. *In* CUTBILL, J. L. (ed.) *Data processing in biology and geology.* Academic Press, London. Pp. 311–320.

ORNA, L. and PETTITT, C. 1980. *Information handling in museums.* Bingley. 190 pp.

PORTER, M. F. 1982. GOS: a package for making catalogues. *Inf. Technol. Res. Dev.* **1**, 113–129.

PRICE, D. 1981. Collections and collectors of note. 39. The Sedgwick Museum, Cambridge. *Geol. Curator*, **3**, 28–35.

RICKARDS, R. B. 1979. The physical basis of palaeontological curating. *Spec. Pap. Palaeont.* **23,** 75–86.
ROBERTS, D. A. and LIGHT, R. B. 1980. Progress in documentation: museum documentation. *J. Docum.* **36,** 42–84.

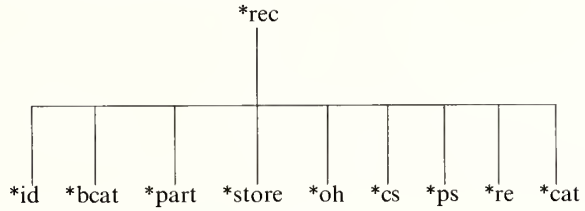
DAVID PRICE
Sedgwick Museum
Department of Earth Sciences
Downing Street
Cambridge CB2 3EQ

Typescript received 6 September 1983
Revised typescript received 14 October 1983

APPENDIX

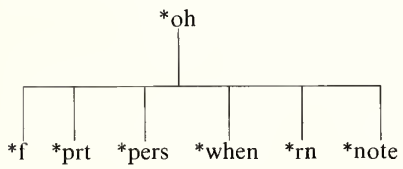
Details of data structure

The SM Format is a much modified version of the 'General Index Format' devised by J. L. Cutbill during the early stages of work at the Sedgwick Museum. It can be visualized as an irregularly branching tree-like structure. The main trunk represents a complete record, and this is progressively divided into increasingly restricted data categories until the terminal branches represent *basic* data fields which, unlike the *group* data fields nearer the trunk, cannot be divided further and contain the actual items of data making up the record, either as strings of characters or as numeric variables. Basic fields are either *keyword* fields containing essential data, or *detail* fields containing data elaborating the keywords. If each data field is represented by its tag, the point where the trunk of the format first divides can be represented as follows:



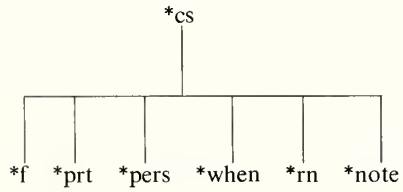
*rec	denotes the complete *record* in SM Format
*id	the record *identity* (= specimen catalogue number)
*bcat	the *broad category* of objects to which the specimen belongs (implicitly 'fossil'; otherwise 'replica', 'artefact', 'rock specimen', or 'inorganic' explicitly stated)
*part	a declaration that the specimen comprises two or more *parts*
*store	the *storage location*
*oh	the *ownership history* (mode of acquisition by the Sedgwick Museum, and previous ownership)
*cs	a *collection statement* (saying by whom the specimen was collected and when)
*ps	a *provenance statement*
*re	a *research event* (anything from an informal identification to the designation of the specimen as a type)
*cat	*cataloguer* information

Of these nine data fields, *oh, *cs, *ps, and *re contain by far the bulk of the data in the record, and are accordingly known as the *main data groups*. They represent major branches of the format which are worth following separately. First, *ownership history*:



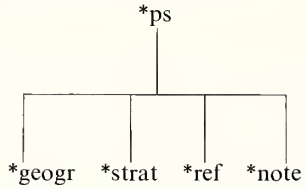
*f	a *function* word (e.g. 'presented', 'exchanged', 'purchased')
*prt	the part(s) of the specimen concerned
*pers	the person or institution concerned
*when	the date or a time period (range of dates)
*rn	the reference number of the specimen in any previous collection
*note	a general detail field

*f and *pers are each made up of keyword and detail fields, *when contains *date denoting a single date (field repeated for a period) and is broken further into separate fields for day, month, and year (each containing a two digit number), with a detail field for terms like 'circa', 'not later than' etc. The *collection statement* field has a similar structure:



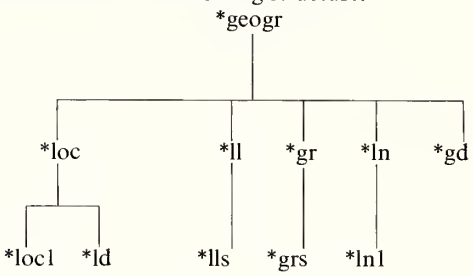
but for this field the function (*f) word 'collected' is automatically generated by the GOS 'display' processor rather than put in as data.

The *provenance statement* has the structure:

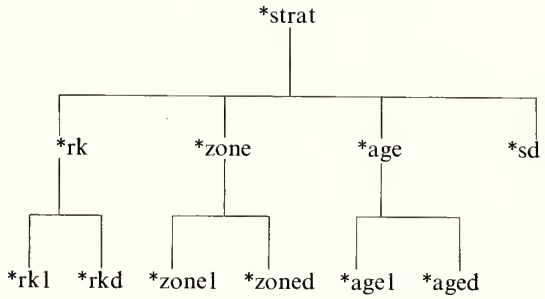


*geogr	contains *geographical* information
*strat	contains *stratigraphical* information
*ref	contains *documentary* information (e.g. bibliographic references)

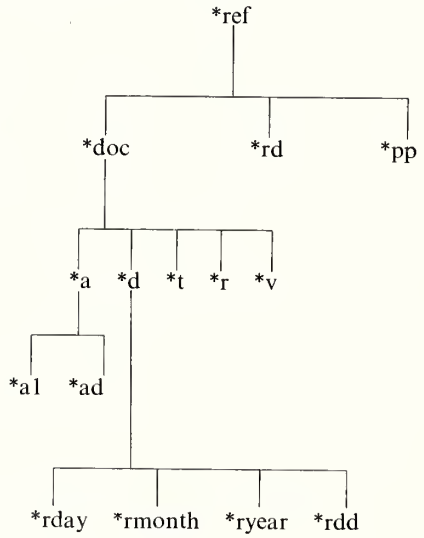
These fields have the following structure:



*loc	place names
*loc1	a specific place name (keyword)
*ld	detail for *loc1 (e.g. 'near', '3 km NW')
*ll	*latitude* and *longitude*
*lls	text for *ll
*gr	*grid reference*
*grs	text for *gr
*ln	locality number
*ln1	text for *ln
*gd	general geographical detail

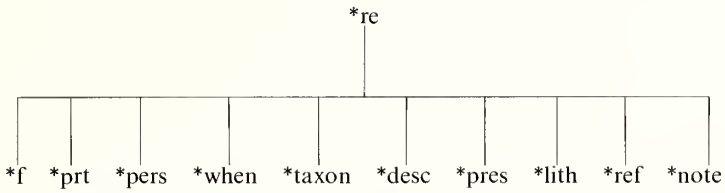


*rk	*lithostratigraphic* information
*rk1	name of rock unit (usually a formational name)
*rkd	detail for *rk1 (e.g. 'basal')
*zone	*biostratigraphic* information
*zone1	name of zone or sub-zone
*zoned	detail for zone1 (e.g. 'topmost')
*age	*chronostratigraphic* information
*age1	name of period, stage, etc.
*aged	detail for *age1
*sd	general stratigraphic detail



*doc	document
*a	*author*
*a1	name of author
*ad	detail for *a1
*d	date
*rday	day (as a two digit number)
*rmonth	month (as a two digit number)
*ryear	year (as a two digit number)
*rdd	any qualifying details for date
*t	title
*r	journal
*v	volume
*rd	reference detail
*pp	page number, plate, and figure information

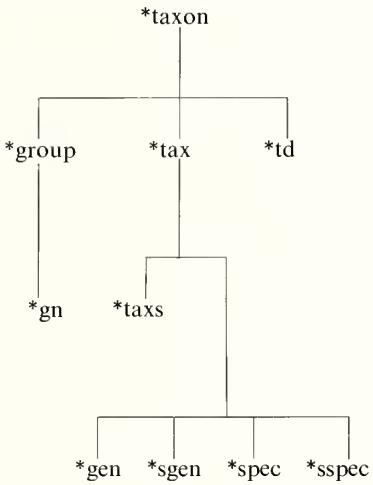
The *research event* is structured thus:



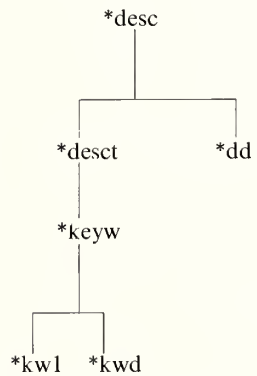
In the *re field, *f contains words such 'described', 'mentioned', 'fig'd', 'holotype', 'paratype', etc.

*taxon	*taxonomic* information	*pres	*preservational* information
*desc	*descriptive* information (morphology)	*lith	*lithological* information (matrix)

*pres and *lith are simply made up of keyword and detail fields; *taxon and *desc are structured as follows:



*group	suprageneric classification	arbitrary museum scheme
*gn	name of suprageneric group	arbitrary museum scheme only
*tax	formal taxonomic nomenclature	
*taxs	full taxonomic name with author(s)	
*gen	*genus*	
*sgen	*subgenus*	
*spec	*species*	
*sspec	*subspecies*	



*desct	descriptive term
*keyw	keyword information
*kwl	actual keyword (e.g. 'cranidium', 'phragmacone')
*kwd	detail for *kwl (e.g. 'incomplete', 'distal end')
*dd	description detail (e.g. 'atypically convex', 'shows doublure')

In designing the SM Format we have tried to produce a data structure which gives much scope for improved specimen documentation; for instance, by incorporating fields for morphological, preservational, and lithological data. The only major data category not catered for is biometric data, and we have considered this to be beyond the scope of a general museum database. It would, of course, be possible to produce specialized databases to contain biometric data for particular groups of organisms for which standardized measurement schemes had been agreed.

The rationale of the SM Format can be illustrated by considering its application to a hypothetical specimen from the *Schloenbachia varians* Zone of the lower Chalk (Cenomanian) of Burwell, Cambridgeshire, collected by J. Bloggs in 1970, identified by him as the external mould of the right valve of the bivalve *Agenus beta* Smith and presented to the Sedgwick Museum in the same year. The specimen is subsequently selected by T. Jones as the holotype of his new species *Agenus nova* Jones, 1975 (*Geol. Mag.* vol. 200, p. 16, pl. 1, fig. 3). Supposing that a catalogue number A1 had been applied to the specimen, and that it had been stored in drawer x.t.114 of the museum, then it would ultimately have a record in the computer-file catalogue with the following structure:

```
*rec
  *id
    *key
      *code
        *elem A
        *elem2 1
        *elem2 0   (this indicates a single
                    specimen)
  *bcat fossil   (implicit—data not actually
                  present)
  *store
    *store1 x.t.114
  *oh
    *f
      *f1 presented
    *pers
      *ps1 Bloggs, J.
    *when
      *date
        *wyear 1970
  *cs
    *f
      *f1 collected   (implicit—data not
                       actually present)
    *pers
      *ps1 Bloggs, J.
    *when
      *date
        *wyear 1970
  *ps
    *geogr
      *loc
        *loc1 Burwell
        *loc1 Cambridge
    *strat
      *rk
        *rk1 Chalk
        *rkd lower
      *zone
        *zone1   /USchloenbachia/N
                 /Uvarians/N Zone†
      *age
        *age1 Cenomanian
        *age1 Cretaceous
*re
  *f
    *f1 identified
  *pers
    *ps1 Bloggs, J.
  *when
    *date
      *wyear 1970
  *taxon
    *group
      *gn Bivalve
    *tax
      *taxs /UAgenus/N /Ubeta/N Smith†
      *gen Agenus
      *spec beta
  *desc
    *desct
      *keyw
        *keyw1 right valve
  *pres
    *pres1 external mould
*re
  *f
    *f1 fig'd
    *f1 holotype
  *taxon
    *group
      *gn Bivalve
    *tax
      *taxs /UAgenus/N /Unova/N Jones†
      *gen Agenus
      *spec nova
*ref
  *doc
    *a
      *a1 Jones, T.
    *d
      *ryear 1975
    *r Geol. Mag.
    *v 200
  *pp p.16, pl.1, fig.3
  *cat
    *catn Price, D.
    *catd 1972
```

(† /U and /N are 'flags' controlling underlining)

This example illustrates an important feature of GOS-based data handling: that fields can be repeated, if necessary many times over. For example, complicated locality descriptions can be broken down into a series of *loc keywords or complicated modes of preservation described by a series of *pres keywords. This feature is of

particular importance in the Sedgwick System in respect to the *re (research event) field. Repetition of this field allows a record to contain all the taxonomic names ever applied to a specimen so that the specimen could be retrieved on any one of them. The hypothetical record with the structure illustrated above would, in fact, generate the following catalogue entry:

A.1 store: x.t.114
Presented, Bloggs, J., 1970.
Collected, Bloggs, J., 1970.
Lower Chalk, *Schloenbachia varians* Zone, Cenomanian, Cretaceous; Burwell, Cambridgeshire.
Identified, Bloggs, J., 1970, as bivalve *Agenus beta* Smith; right valve, external mould.
Fig'd, Holotype, Jones, T., 1970, Geol. Mag., 200, p. 16, pl. 1, fig. 3, as bivalve *Agenus nova* Jones.
[Catalogued Price, D./1972]

With respect to the Information Retrieval System, this hypothetical record would generate the following index terms (from the fields listed on p. 397 above):

'I'	'Rchalk'	'Krightvalve'
'Sx.'	'Qcenomanian'	'Pexternalmould'
'Sx.t.'	'Qcretaceous'	'Ffig'd'
'Sx.t.114'	'Gbivalve'	'Fholotype'
'Dbloggs,j.'	'Tagenus'	'Ajones,t.1970'
'Lburwell'	'Ubeta'	
'Lcambridgeshire'	'Unova'	